Transfer Learning from Simulated to Real Scenes for Monocular 3D Object Detection

Sondos Mohamed*[1], Walter Zimmer*[2], Ross Greer[3], Ahmed Alaaeldin Ghita[2], Modesto Castrillón-Santana[4], Mohan Trivedi[5], Alois Knoll[2], Salvatore Mario Carta[1], and Mirko Marras[1]

[1] University of Cagliari, Cagliari, Italy
{sondos.mohamed,salvatore,mirko.marras}@unica.it
[2] Technical University of Munich, Munich, Germany
{walter.zimmer,ahmed.ghita,k}@tum.de
[3] University of California, Merced, USA
rossgreer@ucmerced.edu
[4] Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain
modesto.castrillon@ulpgc.es
[5] University of California, San Diego, USA
mtrivedi@ucsd.edu

Abstract. Accurately detecting 3D objects from monocular images in dynamic roadside scenarios remains a challenging problem due to varying camera perspectives and unpredictable scene conditions. This paper introduces a two-stage training strategy to address these challenges. Our approach initially trains a model on the large-scale synthetic dataset, *RoadSense3D*, which offers a diverse range of scenarios for robust feature learning. Subsequently, we fine-tune the model on a combination of real-world datasets to enhance its adaptability to practical conditions. Experimental results of the *Cube R-CNN* model on challenging public benchmarks show a remarkable improvement in detection performance, with a mean average precision rising from 0.26 to 12.76 on the *TUM Traffic A9 Highway* dataset and from 2.09 to 6.60 on the *DAIR-V2X-I* dataset, when performing transfer learning. Code, data, and qualitative video results are available on the project website: https://roadsense3d.github.io.

Keywords: Monocular 3D Object Detection, Intelligent Transportation Systems, Intelligent Vehicles, Synthetic Data, Transfer Learning.

1 Introduction

The development of smart cities has become increasingly crucial as urban areas expand and face complex challenges in traffic management and safety. Intersections, for example, are responsible for 40% of major injuries in Canada

* Equal contribution

and vehicle crashes in the United States [5]. Integrating advanced technologies, such as cameras, into monitoring systems is central to the smart city concept. In China alone, around 200 million outdoor cameras are deployed in their Skynet project [42]. While LiDAR and radar are also utilized for traffic monitoring [63, 67], cameras present a cost-effective solution with an extensive range of perception, making them more widely deployable and affordable.

Accurate detection of objects within monocular camera images is paramount for facilitating intelligent monitoring and effective decision-making [1]. Recent advancements in deep learning have fueled a growing interest in 2D/3D object detection approaches. Traditional one-step and two-step 2D object detection methods which predominantly analyze pixel-level information [4, 18, 23, 24, 27, 29, 38, 39, 59], and, more recently, anchor-free and transformer methods, have been involved in traffic monitoring and applications [14–17]. However, methods that provide only 2D detections are limited in providing precise real-world distance measurements between objects as well as the egocentric object. This limitation underscores the necessity for a more comprehensive understanding of the scene and the development of advanced 3D object detection capabilities. Recently, there has been growing interest in infrastructure 3D datasets [9, 52, 58, 62, 65, 66].

Recent monocular 3D models have shown impressive results. However, generalization remains a significant challenge, and most models are domain-specific [7, 35]. Exposing these models to a diverse spectrum of datasets with varying factors can enhance their robustness. Models like *Cube R-CNN* [2] and *UniMode* [26], trained on a wide range of indoor and outdoor datasets, exemplify this approach. Despite their success, these models still face challenges in unfamiliar environments, e.g., roadside scenarios [7]. Another work, *MonoUNI* [21], integrates vehicle and infrastructure data, which adds long-range perception capability. It is evaluated on five benchmarks: *Rope3D* [52], *DAIR-V2X-I*, *KITTI* [12], *Waymo* [46], and *nuScenes* [3]. However, separate training for the vehicle and infrastructure domains is still necessary, and a hybrid training that combines both domains is not yet possible. Also, despite the differences between *DAIR-V2X-I* [55] and *Rope3D* [6] as roadside datasets, they share similarities in their view. On the other hand, the model requires calibration information during the inference, which is often lacking in roadside infrastructure cameras. Consequently, there is a demand for monocular 3D models with zero-shot capability to produce an object's 3D position, size, and orientation (9 attributes per object).

While these models show high accuracy under typical (driving) conditions, their performance degrades significantly when encountering roadside scenarios, such as vehicles that are tilted or overturned due to an accident, primarily due to the limitations in the data annotation process. Specifically, most autonomous driving models predominantly rely on the yaw angle [13], often neglecting the roll and pitch angles because they are zeros. However, these angles are crucial in accurately detecting objects at slight elevations, such as in roadside scenarios. To address limitations, in this paper, we conduct comprehensive transfer learning experiments using the *Cube R-CNN* model, transitioning from synthetic datasets such as *RoadSense3D* [7] to real-world datasets like *TUM Traffic*

A9 Highway (*TUMTraf-A9*) [9] and *DAIR-V2X-I* [55]. In these experiments, we incorporate pitch and roll into both the training and testing phases. The real-world datasets are collected from multiple cities, each with distinct infrastructure configurations, ensuring that the model is exposed to a wide range of urban environments for improved generalization. Through extensive evaluation across these three real-world datasets, we demonstrate that transfer learning improves the $3D\ mAP$ results from 0.26 to 12.76 on the *TUMTraf-A9* dataset and from 2.09 to 6.60 on the *DAIR-V2X-I* dataset, when transitioning from simulated to real scenes. We provide model code, datasets, and qualitative video results on our project website: https://roadsense3d.github.io.

2 Related Work

Our work builds upon prior research focusing on data collection using monocular cameras, as well as methods for detecting 3D bounding boxes from these images.

2.1 Datasets for 3D Monocular Object Detection

Concerted efforts have been made to collect datasets that include 3D bounding box annotations (Table 1). The main factors characterizing these datasets are the domain (vehicle, roadside, or other infrastructure positions), the type (real or synthetic data), and the image characteristics and quantity.

Several datasets are central to the vehicle view domain [3,8,12,19,34,36,37,46]. Among them, *KITTI* [12] is a pioneering and widely utilized dataset that provides benchmarks for various vision-related tasks, including 3D object detection and localization. Datasets such as *nuScenes* [3], *Argoverse* [8], and *Waymo Open* [46] have covered around 1,000 driving scenes, with the latter distinguishing itself by offering the largest number of 3D bounding boxes. *ONCE* provides the largest number of frames. *Argoverse* [8] is notable for including HD maps. Significant advancements in 360-degree camera viewpoints are demonstrated by datasets like *H3D* [36], *nuScenes* [3], *ONCE* [34], and *Argoverse* [8]. The *A*3D* [37] focuses on the high object density and heavy occlusions. Additionally, *ApolloScape* provides image-based 3D instance segmentation and includes object tracking. Despite the multimodality and diverse tasks offered by these datasets, they still focus on cameras mounted on the car, making them more susceptible to obstacles and short-term events. Therefore, other datasets are targeting long-term prediction. *Ko-PER* [45] is one of the pioneering infrastructure datasets, providing 3D object detection data from a permanent intersection monitoring system. Datasets such as *TUMTraf Intersection* [9], *DAIR-V2X-I* [55], *V2X-Seq* [56], and *Rope3D* [52] address various intersection scenarios. *DAIR-V2X-I*, *V2X-Seq*, and recently, *TUMTraf V2X* [65,66] provide views from the vehicle and the roadside infrastructure. *V2X-Seq* offers data for sequential perception which is derived from DAIR-V2X and trajectory forecasting, while *TUMTraf-A9* [9] covers unsequential highway scenarios. Furthermore, BoxCars [44] provides a large-scale intersection dataset as 2D object projections of 3D boxes.

Table 1: Comparison of existing publicly available datasets for the vehicle and roadside infrastructure domains, including their year of release, domain, type, labeling range, RGB resolution, number of RGB images, number of 3D boxes, presence of rain/night data, and availability to the public.

Dataset	Year	Domain	Type	Range	Resolution	Images	3D Boxes	Rain/ Night	Public
KITTI [12]	2013	Vehicle	Real	70m	1392x512	15K	80K	No/No	Yes
KoPER [45]	2014	Roadside	Real	-	656x494	-	-	No/No	Yes
Apolloscape [19]	2018	Vehicle	Real	420m	**3384x2710**	144K	70K	No/Yes	Yes
BoxCars [44]	2018	Roadside	Real	-	128x128	116K	116K	No/No	Yes
nuScenes [3]	2019	Vehicle	Real	75m	1600x900	1.4M	1.4M	Yes/Yes	Yes
Argoverse [8]	2019	Vehicle	Real	200m	1920x1200	22K	993K	Yes/Yes	Yes
H3D [36]	2019	Vehicle	Real	100m	1920x1200	27.7K	1M	No/No	Yes
A*3D [37]	2020	Vehicle	Real	100m	2048x1536	39K	230K	Yes/Yes	Yes
Waymo Open [46]	2020	Vehicle	Real	75m	1920x1080	230K	**12M**	Yes/Yes	Yes
DAIR-V2X-I [55]	2021	Vehicle/Other	Real	200m	1920x1080	71K	1.2M	-/Yes	Yes
BAAI-VANJEE [10]	2021	Roadside	Real	-	1920x1080	5K	74K	Yes/Yes	No
ONCE [34]	2021	Vehicle	Real	200m	1920x1080	**7M**	417K	Yes/Yes	Yes
Rope3D [52]	2022	Roadside	Real	200m	1920x1200	50K	1.5M	Yes/Yes	No
TUMTraf-A9 [9]	2022	Roadside	Real	**700m**	1920x1200	1k	15k	Yes/Yes	Yes
RoadSense3D [56]	2023	Roadside	Synthetic	200	1920x1080	1.4M	9M	Yes/Yes	Yes
V2X-Seq [56]	2023	Vehicle/Other	Real	200	1920x1080	15k	10.45k	Yes/Yes	Yes
TUMTraf Intersection [9]	2023	Roadside	Real	120m	1920x1200	4.8k	62.4k	Yes/Yes	Yes
TUMTraf V2X [65,66]	2024	Vehicle/Roadside	Real	200m	1920x1200	5k	30k	Yes/Yes	Yes
TUMTraf Vehicle [65,66]	2024	Vehicle	Real	200m	1920x1200	1k	30k	Yes/Yes	Yes
TUMTraf Synthetic [58]	2024	Roadside	Synthetic	200m	1920x1200	24k	240k	Yes/Yes	Yes

Besides real data, there are several simulated roadside datasets. *TUMTraf Synthetic* [58] is built using the *CARLA* simulator [11] and follows the *KITTI* format [12]. The camera position and orientation are automatically varied for each frame. It includes diverse weather conditions and provides annotations for ten object classes. The dataset also includes ground truth data for semantic segmentation, depth maps, and RGB images. On the other hand, *RoadSense3D* [6] is composed of 35 intersection areas collected from 7 *CARLA* towns. It is considered the largest roadside dataset in terms of the number of images and 3D bounding boxes, featuring more than 9M 3D bounding boxes and approximately 1.4M frames and considering pitch angles.

The literature shows that real-world datasets including a roadside view may not be large enough to train solid models, while synthetic roadside datasets, although larger, introduce a domain gap when applied to real-world scenarios. Our intuition in this work is to leverage synthetic roadside datasets for initial training, to learn foundational features, and then fine-tune the model on smaller real-world roadside datasets for domain-specific accuracy.

2.2 Methods for Monocular 3D Object Detection

Given the abundance of datasets, several methods have been proposed in the literature to address the 3D bounding box detection task (Table 2). Such methods can be categorized into two main classes: those utilizing 2D features, referred to as result lifting methods, and those leveraging 3D features, which encompass both feature lifting and data lifting methods. [32].

Data lifting methods convert the entire 2D image into a 3D representation. A notable example is the pseudo-LiDAR method, which generates point cloud data from images before applying a LiDAR-based model for detection [53]. Although data lifting methods yield promising results, they are computationally intensive. Consequently, methods with lower computational requirements have been developed, which fall into two categories: result lifting and feature lifting. Result lifting methods transform 2D detections into 3D by estimating depth to recover the corresponding 3D location from image points [2]. Feature lifting methods extract 2D features, lift them into 3D space, and then predict 3D objects by transforming image features into 3D voxel grids or orthographic features [40]. DETR3D [49], MonoDLE [33], GUPNet [31], MonoUNI [21], and Cube R-CNN [2] are examples of result lifting methods. These methods estimate depth from 2D features to recover the corresponding 3D locations. On the other hand, ImVoxelNet [41] and UniMODE [26] are classified as feature lifting methods, where 2D features are lifted into 3D space for object prediction.

Monocular 3D object detection methods can be further categorized by their application context. Methods designed for vehicle view applications operate in dynamic environments, managing varying speeds and rapidly changing perspectives, which necessitates rapid adaptation for accurate depth estimation [20–22, 25, 26, 28, 30, 31, 33, 41, 43, 47–49, 51, 57, 60]. Roadside view models, in contrast, are deployed in static environments and optimized for long-range detection, typically utilizing elevated cameras with wide fields of view [50, 61]. Indoor models must handle shorter focal lengths and significant pitch and yaw variations due to the complex orientations of objects [2, 26].

Despite their application context, several methods can be readily adapted to meet the requirements of roadside infrastructure views, the focus of our study in this paper, as they can extract features over broad ranges and handle different camera configurations.. Specifically, MonoUNI [21] addresses challenges in both roadside infrastructure and vehicle domains by introducing normalized depth to mitigate pitch angle and focal length variations. *ImVoxelNet* [41] generates voxel representations and employs distinct heads for indoor and outdoor environments, accommodating both multi-view and monocular inputs. *Cube R-CNN* [2] extends *Faster R-CNN* with 3D branches, utilizing virtual depth and mesh representations for robust object detection across indoor and outdoor scenes. *UniMODE* [26] adopts a two-stage architecture, addressing grid size challenges and computes loss based on dataset-specific metrics to handle label inconsistencies.

3 Methodology

In this section, we first mathematically formulate the task of 3D object detection from monocular cameras. Next, we introduce the model selection process according to the roadside scenario. We then describe initial training with synthetic data, which involves detailing the dataset and the method for training the model from scratch. We then elaborate on the fine-tuning phase, discussing the selected real-world datasets and the technical aspects of the fine-tuning process.

Table 2: Monocular 3D object detection methods sorted by publication year and method name. The scenario types include vehicle (V), roadside (R), and indoor (I).

Method	Venue & Year	Type V	R	I	Short Description
DETR3D [49]	PMLR 2021	✓			Transformer-based, direct 3D bounding box prediction from images.
GUPNet [31]	ICCV 2021	✓			Geometry uncertainty propagation, geometry-aware.
ImVoxelNet [41]	WACV 2021	✓		✓	Voxel-based, multi-view fusion for improved 3D understanding.
MonoDet [43]	ICCV 2021	✓			Single-stage, anchor-based detection with depth-aware module.
MonoDLE [33]	CVPR 2021	✓			Efficient depth estimation, improved localization with light-weight architecture.
MonoEF [60]	TPAMI 2021	✓			Edge fusion, enhanced depth and boundary prediction.
MonoFlex [57]	CVPR 2021	✓			Anchor-free, flexible regression head, uncertainty-based keypoint estimation.
MonoXiver [48]	ICCVW 2021	✓			Single-frame, uncertainty-guided feature refinement.
PGD [47]	PMLR 2021	✓			Probabilistic depth modeling, geometry-aware detection.
CIE [51]	CoRR 2022	✓			Contextual information extraction, multi-view consistency.
DEVIANT [22]	ECCV 2022	✓			Adversarial training, domain adaptation for robust detection.
MonoCon [30]	AAAI 2022	✓			Anchor-free, context enhancement, auxiliary task integration.
MonoDDE [25]	CVPR 2022	✓			Diversity-driven ensemble, improved generalization across domains.
BEVHeight [50]	CVPR 2023		✓		BEV-based, height-guided feature extraction for roadside detection.
Cube R-CNN [2]	CVPR 2023	✓		✓	3D bounding box regression, multi-modal fusion for indoor scenes.
MonoDet3D [61]	IV 2023		✓		3D object detection for roadside infrastructure sensors.
MonoUNI [21]	NeurIPS 2023	✓	✓		Unified architecture for vehicle and roadside detection, domain adaptation.
Far3D [20]	AAAI 2024	✓			Far-field detection, enhanced distance-aware features for long-range perception.
RayDN [28]	ECCV 2024	✓			Ray-based depth estimation, novel loss function for accurate localization.
UniMODE [26]	CVPR 2024	✓		✓	Unified model for vehicle and indoor detection, multi-domain training.

3.1 Problem Definition

Image-based 3D object detection involves determining the position and shape of objects in three-dimensional space based on a two-dimensional image captured by a camera. To address this, we aim to learn a function, parametrized by θ, that maps a 2D RGB image $i \in \mathcal{I}$, where $\mathcal{I} \subset \mathbb{R}^{H \times W \times 3}$ represents the set of images with height H, width W and corresponding camera parameters, to a set of 3D object attributes. Specifically, for each image i, the model outputs attributes for each detected object j: category $c_{i,j}$, 3D position coordinates $(x_{i,j}, y_{i,j}, z_{i,j})$, dimensions $(h_{i,j}, w_{i,j}, l_{i,j})$, and yaw-pitch-roll orientation angles $(\vartheta_{i,j}, \phi_{i,j}, \psi_{i,j})$. This process can be formalized as:

$$f_\theta(i) \rightarrow \{(c_{i,j}, (x_{i,j}, y_{i,j}, z_{i,j}), (h_{i,j}, w_{i,j}, l_{i,j}), (\vartheta_{i,j}, \phi_{i,j}, \psi_{i,j})) \mid j = 1, \ldots, N_i\} \quad (1)$$

where N_i denotes the number of objects detected in the image i.

To learn a function for yielding these 3D object attributes, we use a dataset $\mathcal{D}$ ($M = |\mathcal{D}|$) consisting of images with annotated 3D bounding boxes. Each entry in the training dataset includes an image i and its ground truth attributes $y_i = \left\{\left(c^*_{i,j}, (x^*_{i,j}, y^*_{i,j}, z^*_{i,j}), (h^*_{i,j}, w^*_{i,j}, l^*_{i,j}), (\vartheta^*_{i,j}, \phi^*_{i,j}, \psi^*_{i,j})\right) \mid j = 1, \ldots, N_i\right\}$, where the asterisks denotes the true values. The training objective is to optimize the function's parameters θ to accurately predict these attributes. In other words, we aim to minimize the loss of the predictions given true annotations. Formally:

$$\hat{\theta} = \arg\min_\theta \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}\left(y_i, f_\theta(i)\right) \quad (2)$$

The typical loss function $\mathcal{L}$ combines several components to address the different aspects of the prediction task, including classification ($\mathcal{L}_{\text{cls}}$), position ($\mathcal{L}_{\text{pos}}$),

dimension ($\mathcal{L}_{\text{dim}}$), and orientation ($\mathcal{L}_{\text{ori}}$). The training process seeks to find the function's parameters θ that minimize this composite loss across the dataset $\mathcal{D}$. By doing so, the function can yield accurate categories, positions, dimensions, and orientations from 2D images, thereby solving the target task.

3.2 Model Selection

To select image-based 3D object detection methods, we conducted an extensive review of top-tier computer vision conferences (e.g., ECCV, CVPR, ICCV) and journals (e.g., IEEE Transactions on Pattern Analysis and Machine Intelligence, T-PAMI) for publications since 2021. Our review focused specifically on models that utilize end-to-end architectures, avoiding those dependent on auxiliary networks for depth extraction. We also prioritized models that have demonstrated strong performance across multiple datasets. Our selection criteria included the models' ability to demonstrate domain adaptability and their applicability across diverse tasks (see Table 2). From this rigorous assessment, we identified four models for further consideration: *MonoUNI* [21], *ImVoxelNet* [41], *Cube R-CNN* [2], and *UniMODE* [26]. Among these, *Cube R-CNN* was chosen for our experiments due to its notable reproducibility and unified training pipeline. Unlike *UniMODE*, which was excluded due to limitations in its training pipeline, *Cube R-CNN* integrates multiple camera coordinate systems and is robust in handling six degrees of freedom in object orientations. In contrast to *ImVoxelNet* and *MonoUNI*, which rely on dataset-specific training strategies, *Cube R-CNN* does not compromise effectiveness across datasets.

From an architectural perspective, *Cube R-CNN* builds upon the Faster R-CNN framework [39], an end-to-end region-based object detection approach. Faster R-CNN employs a backbone network, typically a convolutional neural network (CNN), to transform the input image into a higher-dimensional feature representation. A Region Proposal Network (RPN) then generates regions of interest (ROIs) that signify potential object candidates within the image. These ROIs are processed by a 2D box head, which uses the backbone's feature map to classify the object and refine the 2D bounding box predictions. A cube head that computes the 3D parameters, including central point projection, depth, scaled dimensions, and object-centered orientation, is applied for each detected object.

3.3 Initial Model Creation

Synthetic Dataset Selection. In our experiments, we used the *RoadSense3D* [6] synthetic dataset comprising over 9 million labeled 3D objects across 1.4 million frames for model training. As detailed in Table 1, this dataset offers a diverse range of scenarios generated from 35 roadside cameras across seven distinct towns in the *CARLA* simulator [11]. Key parameters include 1920x1080 image resolution, 40,448 frames per position, camera pitch angles ranging from -25° to -45°, a detection range of 150 meters, and a 120° field of view. To simulate realistic conditions, the dataset incorporates variations in weather (sunny, cloudy, foggy) and time of day (day/night).

Fig. 1: Qualitative Results on the Synthetic *RoadSense3D* Test Set. We show 3D box detections of the *Cube R-CNN* model in the class-specific colors during different lighting and weather conditions.

We selected the *RoadSense3D* dataset since there is a lack of large real-world datasets that provide sufficient data from roadside environments. Indeed, although datasets like *Rope3D* [52] include millions of real-world images, they are not publicly available. Training models from scratch requires a substantial amount of data, which is not readily accessible from the real world in this domain since labeling is costly and often requires manual labeling effort [64]. Given this, among artificial datasets, *RoadSense3D* is the largest synthetic dataset available and offers comprehensive coverage of various factors.

Training from Scratch. We trained *Cube R-CNN* on the *RoadSense3D* synthetic dataset to address object pose variability. The dataset is sequential, with multiple images containing the same objects in different positions. This strategy allowed the model to encounter various objects, locations, and occlusion scenarios. According to the original paper, we split the dataset into training, validation, and testing sets. The model was trained for 250,000 iterations on a single GPU, using a batch size of 4 and a learning rate of 0.0025. We utilized the Stochastic Gradient Descent (SGD) solver, and the model was evaluated every 10,000 iterations. These hyperparameters were chosen to address inconsistencies observed in the original *Cube R-CNN* model's training. Figure 1 presents qualitative results.

3.4 Pretrained Model Transfer

Real-World Datasets Selection. To enable the *Cube R-CNN* model to generalize to real-world scenarios, we fine-tuned it on several diverse datasets that vary in camera setups and environmental conditions. Due to their public availability and size, we selected the *TUMTraf-A9* and *DAIR-V2X-I* datasets. The *TUMTraf-A9* dataset, a subset of the *TUMTraf* dataset family, captures complex highway scenarios in Munich, Germany, under diverse weather and lighting conditions. It features 1,000 labeled frames with 15,000 3D bounding boxes and track IDs and includes data from both LiDAR and multi-view cameras, using

16mm and 50mm focal lengths to monitor traffic across 11 lanes. The *DAIR-V2X-I* dataset, collected in China, represents a large-scale vehicle-infrastructure cooperative autonomous driving dataset, offering over 71,000 LiDAR and camera frames from both infrastructure and vehicle perspectives.

Pre-processing was needed for the *DAIR-V2X-I* dataset, which provides object annotations in the LiDAR coordinate system. Since *Cube R-CNN* operates in the camera coordinate system, we first transformed the annotations from the LiDAR coordinate system to the camera coordinate system. In the LiDAR system, annotations specify the object's location, dimensions, and yaw rotation across seven degrees of freedom. When projecting these annotations to the camera's view, it is essential to account for the object's rotation relative to the camera. This involves projecting the annotations around all three axes. The transformation process follows the equation $\begin{bmatrix} x\ y\ w \end{bmatrix} = K \cdot T \cdot \begin{bmatrix} X\ Y\ Z\ 1 \end{bmatrix}$, where the intrinsic matrix K and the extrinsic matrix T are used to project the 3D point (X, Y, Z) in the world coordinate system to the 2D point (x, y) in the camera coordinate system. The variable w serves as a scale factor in this transformation, ensuring the proper projection of points into the camera view. We then utilized a 60/40 ratio for training and testing. For the test, we used sequential perception from *V2X-Seq* [56]. For *TUMTraf-A9*, we applied the same 60/40 ratio for training and testing, using data from both the north and south cameras and incorporating both small and large focal lengths.

Fine-tuning. The model was initially trained on the *RoadSense3D* dataset for 250,000 iterations. Subsequently, we began training with a reduced learning rate of $\alpha = 0.0025$, utilizing the DLA34 architecture [54] for feature extraction. Fine-tuning was focused on the head component of the model, which includes both the 2D head and the 3D cube head. The model was further trained for an additional 850,000 iterations.

4 Experimental Results

In this section, we present the results obtained from our transfer learning experiments. First, we examine the impact of transferring from synthetic to real data in a single step, specifically from *RoadSense3D* to *TUMTraf-A9* and *DAIR-V2X-I* separately. Next, we explore the transition from a synthetic dataset to real-world datasets gradually, moving from *RoadSense3D* to *DAIR-V2X-I* and then to *TUMTraf-A9*. The model's performance was assessed on the test set of each real-world dataset using the 3D mean Average Precision (mAP_{3D}) under a certain Intersection over Union (IoU) threshold, measuring both detection and localization precision.

4.1 Single-Step Dataset Transfer

In our initial analysis, we investigated the performance of direct transfer learning. To this end, we trained a *Cube R-CNN* model on the full *RoadSense3D* dataset

Table 3: Single-Step Dataset Transfer on *TUMTraf-A9*. We report the 3D mean average precision across the easy, moderate, and hard difficulty levels. Transfer learning involves pre-training on the synthetic *RoadSense3D* dataset and fine-tuning on the real-world *TUMTraf-A9* dataset.

Architecture	Pre-Train Set	Fine-Tuning Set	Evaluation Set	Difficulty Level		
				Easy	Moderate	Hard
Cube R-CNN	TUMTraf-A9 Train	-	TUMTraf-A9 Test	0.26	0.26	0.26
Cube R-CNN	RoadSense3D Train	TUMTraf-A9 Train	TUMTraf-A9 Test	**12.76**	**12.76**	**12.76**

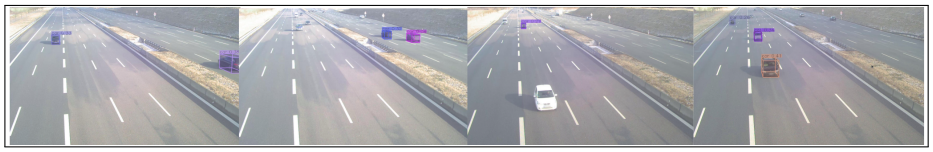
Table 4: Single-Step Dataset Transfer on *DAIR-V2X-I*. We report the 3D mean average precision across the easy, moderate, and hard difficulty levels. Transfer learning involves pre-training on the synthetic *RoadSense3D* dataset and fine-tuning on the real-world *DAIR-V2X-I* dataset.

Architecture	Pre-Train Set	Fine-Tuning Set	Evaluation Set	Difficulty Level		
				Easy	Moderate	Hard
Cube R-CNN	DAIR-V2X-I Train	-	DAIR-V2X-I Test	2.09	2.62	2.61
Cube R-CNN	RoadSense3D Train	DAIR-V2X-I Train	DAIR-V2X-I Test	**6.60**	**8.60**	**8.65**

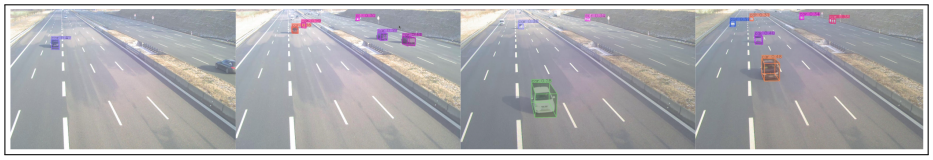
and subsequently fine-tuned it separately on the training sets of *DAIR-V2X-I* and *TUMTraf-A9*, resulting in two fine-tuned models. To assess the impact of transfer learning, we also trained independent *Cube R-CNN* models from scratch on the training sets of *DAIR-V2X-I* and *TUMTraf-A9*, resulting in two additional models for comparison. We then evaluated the transferability of these models on the test sets of the corresponding real-world dataset, *DAIR-V2X-I* and *TUMTraf-A9*, respectively. Tables 3 and 4 collect the mAP_{3D} across difficulty levels for the two model instances when evaluated on the test part of the respective real-world dataset, respectively.

For the *TUMTraf-A9* test set, the model pre-trained on *RoadSense3D* and fine-tuned on *TUMTraf-A9* achieved a mAP_{3D} of 12.76 across easy, moderate, and hard difficulty levels. This represents an increase of 4,808% compared to the mAP_{3D} of 0.26 for the model trained from scratch on *TUMTraf-A9*. Notably, the values for easy, moderate, and hard difficulty levels are identical (12.76). This occurs because the *TUMTraf-A9* test set lacks occlusions. Without them, which typically increase the detection task complexity, the model does not face additional challenges across difficulty levels. The gains observed quantitatively can be better appreciated in Figure 2, which provides qualitative examples of detections for both the considered models.

For the *DAIR-V2X-I* test set, the model fine-tuned with transfer learning achieved mAP_{3D} scores of 6.60 (easy), 8.60 (moderate), and 8.65 (hard), compared to 2.09, 2.62, and 2.61 for the model trained from scratch. These results indicate performance improvements of 215.8% (easy), 228.6% (moderate), and 231.4% (hard), showing the effectiveness of transfer learning across all difficulty levels. The varying gains across difficulty levels also suggest that the benefits



(a) Model trained from scratch on *TUMTraf-A9*.



(b) Model pre-trained on *RoadSense3D* and fine-tuned on *TUMTraf-A9*.

Fig. 2: Qualitative Results of Cube R-CNN on the *TUMTraf-A9* Test Set. Comparison between the *Cube R-CNN* model trained from scratch on *TUMTraf-A9* (top) and model trained on *RoadSense3D* and fine-tuned on *TUMTraf-A9* (bottom).

Table 5: Multi-Step Dataset Transfer on *TUMTraf-A9*. We report the 3D mean average precision across the easy, moderate, and hard difficulty levels. Transfer learning involves training on synthetic *RoadSense3D*, then tuning on *DAIR-V2X-I*, and finally on *TUMTraf-A9*.

Architecture	Pre-Train Set	Fine-Tuning Set	Evaluation Set	Difficulty Level		
				Easy	Moderate	Hard
Cube R-CNN	TUMTraf-A9 Train	-	TUMTraf-A9 Test	0.26	0.26	0.26
Cube R-CNN	RoadSense3D Train	DAIR-V2X Train → TUMTraf-A9 Train	TUMTraf-A9 Test	**6.26**	**6.26**	**6.26**

of transfer learning are more pronounced in moderate and hard scenarios, likely due to the increased complexity of these cases.

In summary, these experiments show that transfer learning from synthetic data can substantially improve real-world performance, particularly when the target dataset is small or has complex scenarios.

4.2 Multi-Step Dataset Transfer

In our second analysis, we investigated the performance of gradual transfer learning. To this end, we used the *Cube R-CNN* model pre-trained on the full *RoadSense3D* dataset and subsequently fine-tuned it first on the larger training set of *DAIR-V2X-I* and then on the smaller training set of *TUMTraf-A9*. To assess the impact of transfer learning, we also trained independent *Cube R-CNN* models from scratch on the training set of *TUMTraf-A9*, resulting in another model for comparison. We then evaluated the transferability of these models on the test set of *TUMTraf-A9*. Table 5 collects the mAP_{3D} across difficulty levels for the two model instances.

Results show that the *Cube R-CNN* model pre-trained on *RoadSense3D* and then fine-tuned sequentially on *DAIR-V2X-I* and *TUMTraf-A9* achieved a 3D mean Average Precision (mAP_{3D}) of 6.26 across all difficulty levels on the

TUMTraf-A9 test set. This represents a substantial increase of 2,308% compared to the 0.26 mAP_{3D} obtained by the model trained from scratch on *TUMTraf-A9*. However, the model fine-tuned directly on *TUMTraf-A9* after pre-training on *RoadSense3D* (Table 3) achieved a higher mAP_{3D} of 12.76 across all difficulty levels. This indicates that while the multi-step transfer learning approach can improve performance, it still falls short of the results obtained through direct fine-tuning on *TUMTraf-A9*. We conjecture that it is caused by potential domain gaps introduced during the intermediate *DAIR-V2X-I* phase. This intermediate step might cause the model to adapt real-world features from *DAIR-V2X-I* that are less optimal for *TUMTraf-A9*, influencing the quality of the subsequent fine-tuning phase and leading to lower overall performance compared to the more direct fine-tuning approach.

In summary, these experiments show that while multi-step transfer learning offers performance improvements, the direct fine-tuning approach tends to be more effective for maximizing performance on specific real-world datasets.

5 Conclusion and Future Work

In this work, we conducted extensive transfer learning experiments using the *Cube R-CNN* model, transitioning from synthetic datasets like *RoadSense3D* to real-world datasets such as *TUMTraf-A9* and *DAIR-V2X-I*. By incorporating pitch and roll into both training and testing phases and evaluating across multiple cities with diverse infrastructure, we demonstrated significant improvements in detection accuracy. Direct transfer learning enhanced the $3D\ mAP$ from 0.26 to 12.76 on the *TUMTraf-A9* dataset and from 2.09 to 6.60 on the *DAIR-V2X-I* dataset, showcasing substantial gains in real-world performance. Our findings indicate that while multi-step transfer learning is beneficial, direct fine-tuning on the target dataset yields superior results. This approach bridges the simulation-to-real gap and paves the way for more robust and adaptable models in intelligent transportation systems. The potential applications extend beyond traffic monitoring to include autonomous driving and smart city infrastructure, where accurate and scalable 3D perception is critical for enhancing safety and efficiency.

Future research will explore adapting additional monocular object detection methods to the existing transfer learning framework, incorporating all yaw, pitch, and roll variations to enhance their adaptability to roadside scenarios. We plan to involve detailed inspections to identify scenarios where transfer learning falls short, aiming to inform the development of novel approaches. Furthermore, integrating active learning and knowledge distillation will be pursued to refine the transfer learning process, focusing on selecting only the most informative examples to let the model adapt. Additionally, incorporating these 3D object detection methods into anomaly detection pipelines for real-world smart city applications will be investigated, with particular emphasis on enhancing accident detection and prevention strategies thanks to more precise object detection.

Acknowledgments

This research was supported by the Federal Ministry of Education and Research in Germany within the project *AUTOtech.agil*, Grant Number: 01IS22088U. Furthermore, this work has been partially supported by the Autonomous Region of Sardinia through Sardegna Ricerche, within the funding call *Aiuti per Progetti di Ricerca e Sviluppo - Settore ICT (2022)*, under the project *SENTINEL: Sustainable intElligeNce-based soluTion for envIronmental and urbaN survEilLance*, Grant Number: G27H23000140002.

References

1. Atzori, A., Barra, S., Carta, S., Fenu, G., Podda, A.S.: Heimdall: an ai-based infrastructure for traffic monitoring and anomalies detection. In: Proc. of the IEEE International Conf. on Pervasive Comp. and Commu., Workshops and Affiliated Events, PERCOM 2021. pp. 154–159. IEEE (2021)
2. Brazil, G., Kumar, A., Straub, J., Ravi, N., Johnson, J., Gkioxari, G.: Omni3d: A large benchmark and model for 3d object detection in the wild. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023. pp. 13154–13164. IEEE (2023)
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
4. Cao, J., Cholakkal, H., Anwer, R.M., Khan, F.S., Pang, Y., Shao, L.: D2det: Towards high quality object detection and instance segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 11482–11491. Computer Vision Foundation / IEEE (2020)
5. Carrillo, J., Waslander, S.L.: Urbannet: Leveraging urban maps for long range 3d object detection. In: Proc. of the 24th IEEE International Intelligent Transportation Systems Conference, ITSC 2021. pp. 3799–3806. IEEE (2021)
6. Carta, S., Castrillón-Santana, M., Marras, M., Mohamed, S., Podda, A.S., Saia, R., Sau, M., Zimmer, W.: RoadSense3d: A framework for roadside monocular 3d object detection. In: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. pp. 452–459. UMAP Adjunct '24, Association for Computing Machinery
7. Carta, S., Santana, M.C., Marras, M., Mohamed, S., Podda, A.S., Saia, R., Sau, M., Zimmer, W.: Roadsense3d: A framework for roadside monocular 3d object detection. In: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP Adjunct 2024, Cagliari, Italy, July 1-4, 2024. ACM (2024)
8. Chang, M., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., Hays, J.: Argoverse: 3d tracking and forecasting with rich maps. Computer Vision Foundation / IEEE (2019)
9. Creß, C., Zimmer, W., Strand, L., Fortkord, M., Dai, S., Lakshminarasimhan, V., Knoll, A.: A9-dataset: Multi-sensor infrastructure-based dataset for mobility research. In: 2022 IEEE Intelligent Vehicles Symposium (IV). pp. 965–970 (2022)

10. Deng, Y., Wang, D., Cao, G., Ma, B., Guan, X., Wang, Y., Liu, J., Fang, Y., Li, J.: BAAI-VANJEE roadside dataset: Towards the connected automated vehicle highway technologies in challenging environments of china. CoRR **abs/2105.14370** (2021)
11. Dosovitskiy, A., Ros, G., Codevilla, F., López, A.M., Koltun, V.: CARLA: an open urban driving simulator. vol. 78, pp. 1–16. PMLR (2017)
12. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012. pp. 3354–3361 (2012)
13. Greer, R., Deo, N., Trivedi, M.: Trajectory prediction in autonomous driving with a lane heading auxiliary loss. IEEE Robotics and Automation Letters **6**(3), 4907–4914 (2021)
14. Greer, R., Desai, S., Rakla, L., Gopalkrishnan, A., Alofi, A., Trivedi, M.: Pedestrian behavior maps for safety advisories: Champ framework and real-world data analysis. In: 2023 IEEE Intelligent Vehicles Symposium (IV). pp. 1–8. IEEE (2023)
15. Greer, R., Gopalkrishnan, A., Deo, N., Rangesh, A., Trivedi, M.: Salient sign detection in safe autonomous driving: Ai which reasons over full visual context. In: 27th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration. No. 23-0333 (2023)
16. Greer, R., Gopalkrishnan, A., Keskar, M., Trivedi, M.M.: Patterns of vehicle lights: Addressing complexities of camera-based vehicle light datasets and metrics. Pattern Recognition Letters **178**, 209–215 (2024)
17. Greer, R., Gopalkrishnan, A., Landgren, J., Rakla, L., Gopalan, A., Trivedi, M.: Robust traffic light detection using salience-sensitive loss: Computational framework and evaluations. In: 2023 IEEE Intelligent Vehicles Symposium (IV). pp. 1–7. IEEE (2023)
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. CoRR **abs/1703.06870** (2017)
19. Huang, X., Cheng, X., Geng, Q., Cao, B., Zhou, D., Wang, P., Lin, Y., Yang, R.: The apolloscape dataset for autonomous driving. In: Proc. of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018. pp. 954–960 (2018)
20. Jiang, X., Li, S., Liu, Y., Wang, S., Jia, F., Wang, T., Han, L., Zhang, X.: Far3d: Expanding the horizon for surround-view 3d object detection. In: Wooldridge, M.J., Dy, J.G., Natarajan, S. (eds.) Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada. pp. 2561–2569. AAAI Press (2024)
21. Jinrang, J., Li, Z., Shi, Y.: MonoUNI: A unified vehicle and infrastructure-side monocular 3d object detection network with sufficient depth clues. vol. 36, pp. 11703–11715
22. Kumar, A., Brazil, G., Corona, E., Parchami, A., Liu, X.: Deviant: Depth equivariant network for monocular 3d object detection. In: European Conference on Computer Vision. pp. 664–683. Springer (2022)
23. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV. Lecture Notes in Computer Science, vol. 11218, pp. 765–781. Springer (2018)

24. Li, Y., Chen, Y., Wang, N., Zhang, Z.: Scale-aware trident networks for object detection. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 6053–6062. IEEE (2019)
25. Li, Z., Qu, Z., Zhou, Y., Liu, J., Wang, H., Jiang, L.: Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2791–2800 (2022)
26. Li, Z., Xu, X., Lim, S., Zhao, H.: UniMODE: Unified monocular 3d object detection. pp. 16561–16570
27. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 2999–3007. IEEE Computer Society (2017)
28. Liu, F., Huang, T., Zhang, Q., Yao, H., Zhang, C., Wan, F., Ye, Q., Zhou, Y.: Ray denoising: Depth-aware hard negative sampling for multi-view 3d object detection. In: 18th European Conference on Computer Vision (ECCV) (2024)
29. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: SSD: single shot multibox detector. In: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I. Lecture Notes in Computer Science, vol. 9905, pp. 21–37. Springer (2016)
30. Liu, X., Xue, N., Wu, T.: Learning auxiliary monocular contexts helps monocular 3d object detection. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. pp. 1810–1818. AAAI Press (2022)
31. Lu, Y., Ma, X., Yang, L., Zhang, T., Liu, Y., Chu, Q., Yan, J., Ouyang, W.: Geometry uncertainty projection network for monocular 3d object detection. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 3091–3101. IEEE (2021)
32. Ma, X., Ouyang, W., Simonelli, A., Ricci, E.: 3d object detection from images for autonomous driving: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **46**(5), 3537–3556 (2024)
33. Ma, X., Zhang, Y., Xu, D., Zhou, D., Yi, S., Li, H., Ouyang, W.: Delving into localization errors for monocular 3d object detection. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. pp. 4721–4730. Computer Vision Foundation / IEEE (2021)
34. Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., Yu, J., Xu, C., Xu, H.: One million scenes for autonomous driving: ONCE dataset (2021)
35. Moon, S., Bae, J., Im, S.: Rotation matters: Generalized monocular 3d object detection for various camera systems. CoRR **abs/2310.05366** (2023)
36. Patil, A., Malla, S., Gang, H., Chen, Y.: The H3D dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In: Proc. of the International Conference on Robotics and Automation, ICRA 2019. pp. 9552–9557. IEEE (2019)
37. Pham, Q.H., Sevestre, P., Pahwa, R.S., Zhan, H., Pang, C.H., Chen, Y., Mustafa, A., Chandrasekhar, V., Lin, J.: A*3d dataset: Towards autonomous driving in challenging environments. In: Proc. of the International Conference in Robotics and Automation, ICRA 2020 (2020)

38. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 779–788. IEEE Computer Society (2016)
39. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. pp. 91–99 (2015)
40. Roddick, T., Kendall, A., Cipolla, R.: Orthographic feature transform for monocular 3d object detection. In: 30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019. p. 285. BMVA Press (2019)
41. Rukhovich, D., Vorontsova, A., Konushin, A.: Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2397–2406 (2022)
42. Seoane, M.F.V., Velasco, C.M.Á.: The chinese surveillance state in latin america? evidence from argentina and ecuador. Inf. Soc. **40**(2), 154–167 (2024)
43. Shi, X., Ye, Q., Chen, X., Chen, C., Chen, Z., Kim, T.: Geometry-based distance decomposition for monocular 3d object detection. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 15152–15161. IEEE (2021)
44. Sochor, J., Špaňhel, J., Herout, A.: Boxcars: Improving fine-grained recognition of vehicles using 3-d bounding boxes in traffic surveillance. IEEE Transactions on Intelligent Transportation Systems **PP**(99), 1–12 (2018)
45. Strigel, E., Meissner, D.A., Seeliger, F., Wilking, B., Dietmayer, K.: The ko-per intersection laserscanner and video dataset. IEEE (2014)
46. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
47. Wang, T., Xinge, Z., Pang, J., Lin, D.: Probabilistic and geometric depth: Detecting objects in perspective. In: Conference on Robot Learning. pp. 1475–1485. PMLR (2022)
48. Wang, T., Zhu, X., Pang, J., Lin, D.: FCOS3D: fully convolutional one-stage monocular 3d object detection. In: Proc. of the IEEE/CVF International Conf. on Computer Vision Workshops, ICCVW 2021. pp. 913–922. IEEE (2021)
49. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: Conference on Robot Learning. pp. 180–191. PMLR (2022)
50. Yang, L., Yu, K., Tang, T., Li, J., Yuan, K., Wang, L., Zhang, X., Chen, P.: Bevheight: A robust framework for vision-based roadside 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21611–21620 (2023)
51. Ye, Q., Jiang, L., Du, Y.: Consistency of implicit and explicit features matters for monocular 3d object detection. CoRR **abs/2207.07933** (2022)
52. Ye, X., Shu, M., Li, H., Shi, Y., Li, Y., Wang, G., Tan, X., Ding, E.: Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022. pp. 21309–21318. IEEE (2022)

53. You, Y., Wang, Y., Chao, W., Garg, D., Pleiss, G., Hariharan, B., Campbell, M.E., Weinberger, K.Q.: Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020)
54. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 2403–2412. Computer Vision Foundation / IEEE Computer Society (2018)
55. Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., Nie, Z.: DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. CoRR **abs/2204.05575** (2022)
56. Yu, H., Yang, W., Ruan, H., Yang, Z., Tang, Y., Gao, X., Hao, X., Shi, Y., Pan, Y., Sun, N., et al.: V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5486–5495 (2023)
57. Zhang, Y., Lu, J., Zhou, J.: Objects are different: Flexible monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3289–3298 (2021)
58. Zhou, X., Fu, D., Zimmer, W., Liu, M., Lakshminarasimhan, V., Strand, L., Knoll, A.C.: Warm-3d: A weakly-supervised sim2real domain adaptation framework for roadside monocular 3d object detection. In: 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC). pp. 1–8
59. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. CoRR **abs/1904.07850** (2019)
60. Zhou, Y., He, Y., Zhu, H., Wang, C., Li, H., Jiang, Q.: Monoef: Extrinsic parameter free monocular 3d object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 10114–10128 (2021)
61. Zimmer, W., Birkner, J., Brucker, M., Tung Nguyen, H., Petrovski, S., Wang, B., Knoll, A.C.: InfraDet3d: Multi-modal 3d object detection based on roadside infrastructure camera and LiDAR sensors. In: 2023 IEEE Intelligent Vehicles Symposium (IV). pp. 1–8. ISSN: 2642-7214
62. Zimmer, W., Creß, C., Nguyen, H.T., Knoll, A.C.: TUMTraf intersection dataset: All you need for urban 3d camera-LiDAR roadside perception. In: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). pp. 1030–1037. ISSN: 2153-0017
63. Zimmer, W., Grabler, M., Knoll, A.: Real-time and robust 3d object detection within road-side lidars using domain adaptation. arXiv preprint arXiv:2204.00132 (2022)
64. Zimmer, W., Rangesh, A., Trivedi, M.: 3d BAT: A semi-automatic, web-based 3d annotation toolbox for full-surround, multi-modal data streams. In: 2019 IEEE Intelligent Vehicles Symposium (IV). pp. 1816–1821. ISSN: 2642-7214
65. Zimmer, W., Wardana, G.A., Sritharan, S., Zhou, X., Song, R., Knoll, A.C.: Tumtraf v2x cooperative perception dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 10 (2024)
66. Zimmer, W., Wardana, G.A., Sritharan, S., Zhou, X., Song, R., Knoll, A.C.: Tumtraf v2x cooperative perception dataset - supplementary material. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 13 (2024)
67. Zimmer, W., Wu, J., Zhou, X., Knoll, A.C.: Real-time and robust 3d object detection with roadside LiDARs. In: Antoniou, C., Busch, F., Rau, A., Hariharan, M.

(eds.) Proceedings of the 12th International Scientific Conference on Mobility and Transport: Mobility Innovations for Growing Megacities. pp. 199–219. Springer Nature